Universidade Federal do Espírito Santo
Centro Tecnológico
Departamento de Informática

Identificação de anomalias em redes a partir de estruturas probabilísticas

Alunos: Athus Cavalini e Thiago Souza
Professor: Rodolfo Villaça

Introdução

O reconhecimento de anomalias em redes, como ataques de negação de serviço (DDoS) é algo cada vez mais necessário, considerando a proporção que o uso da Internet têm alcançado e o quanto isso impacta em nosso dia a dia.

Ao mesmo tempo, o altíssimo (e cada vez maior) fluxo de informações faz com que seja difícil realizar análises sobre esses dados, especialmente considerando as limitações dos dispositivos de roteamento.

Nesse contexto, surge a possibilidade da utilização de estruturas probabilísticas para armazenar informações de tráfego e serem utilizadas por algoritmos de aprendizado de máquinas para identificação de anomalias.

Estruturas Probabilísticas

As Estruturas Probabilísticas são estruturas de dados que visam o armazenamento de informações de modo a reduzir o espaço de memória utilizado e a complexidade dos dados representados.

O termo "probabilística" se dá pelo fato de que, ao contrário das estruturas determinísticas, elas utilizam uma indexação baseada em hashing.

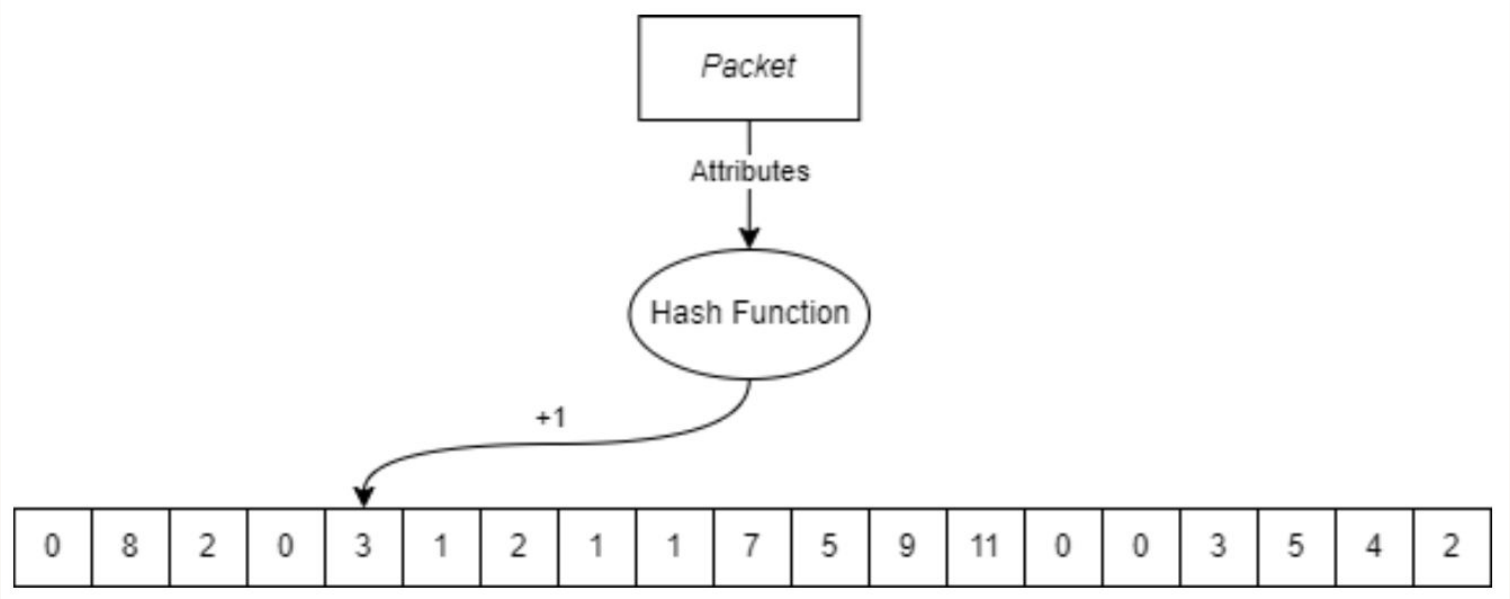
Estruturas Probabilísticas: Counter-Min

O Counter-Min é uma estrutura muito utilizada para contagem de determinados elementos em um fluxo de dados.

O método consiste na utilização de uma matriz (ou um conjunto de listas) cujas posições são preenchidas a partir da aplicação de um hash às informações de entrada.

No contexto de análise de redes, a função de hash se baseia em dados obtidos dos pacotes do fluxo de dados, ou seja, que trafegam pelo roteador ou outro equipamento onde se está montando a estrutura.

Estruturas Probabilísticas: Counter-Min



(PICOLI, 2022)

Metodologia

- Desenvolver uma topologia de rede no Mininet Wifi e simular um tráfego de normalidade entre dispositivos;

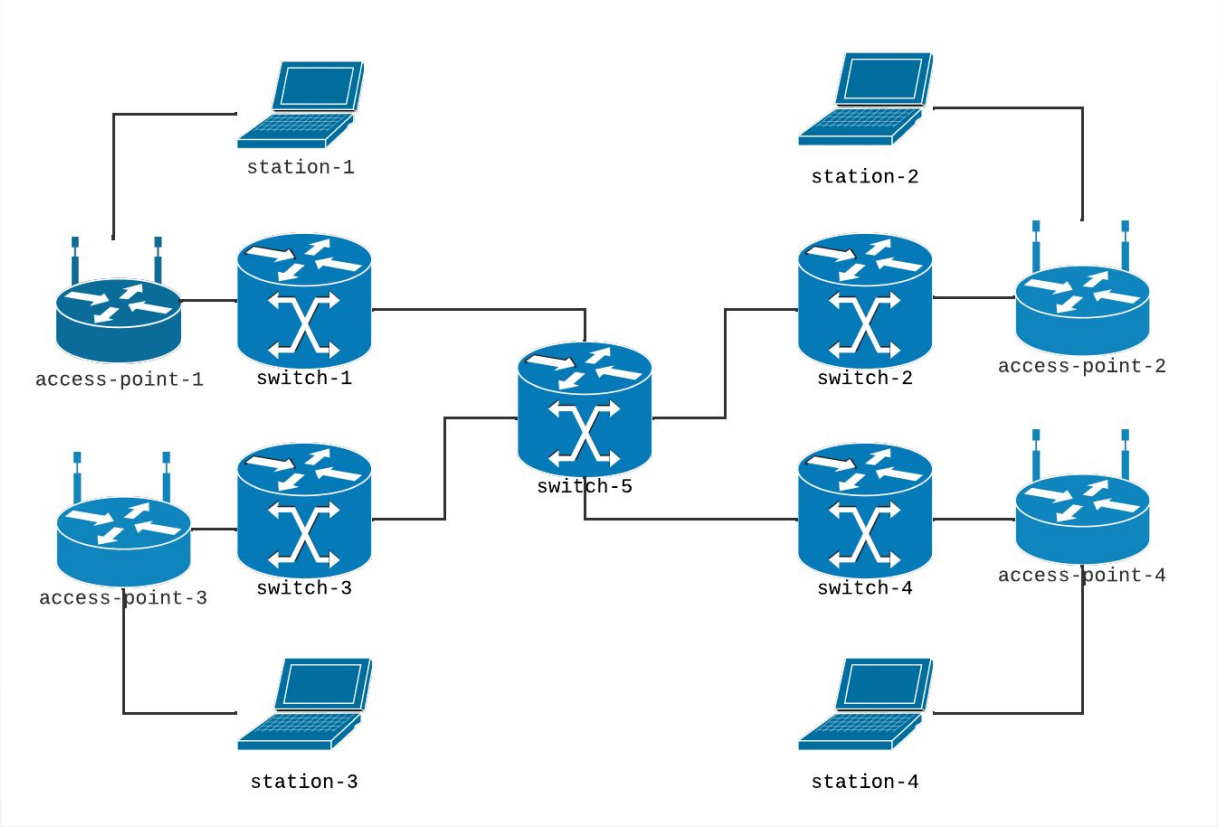
- Adicionar anomalia ao tráfego (elephant flows);

- Salvar o pacotes da rede e utilizá-los para gerar um conjunto de estruturas probabilísticas;

- Extrair métricas das estruturas probabilísticas (média, não-zeros, etc.);

- Aplicar Aprendizado de Máquina às métricas extraídas a fim de identificar intervalos de tempo com tráfego fora da normalidade.

Topologia desenvolvida



Definição do fluxo e geração dos datasets

Para a geração do dataset, utilizamos a ferramenta hping3 para simular o fluxo de dados na rede. Dessa forma, foram definidos os padrões de:

`Normalidade:` 2 a 8 fluxos simultâneos, entre origens e destinos aleatórios, com 1 a 2 pacotes por segundo de 20 a 500 bytes.

`Anomalia:` além do fluxo de normalidade, um fluxo entre origem e destino específicos, com 10 pacotes por segundo de 500 a 1400 bytes cada.

Durante a execução do experimento, os dados eram salvos no formato PCAP a partir do fluxo interceptado no Switch 5.

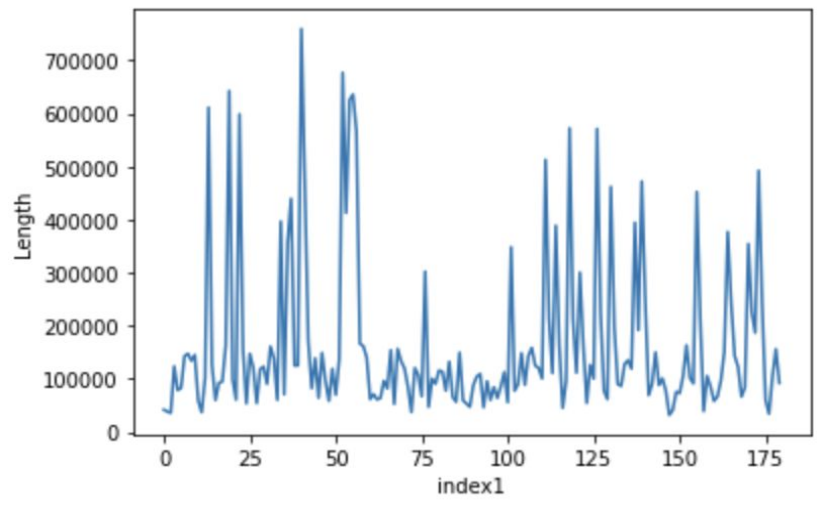
Definição do fluxo e geração dos datasets

O script para a geração do dataset permite configurar o tempo total do experimento e o número de fluxos normais e anômalos disparados neste período. Assim, divide o período em intervalos iguais para distribuir aleatoriamente o número de fluxos definidos.

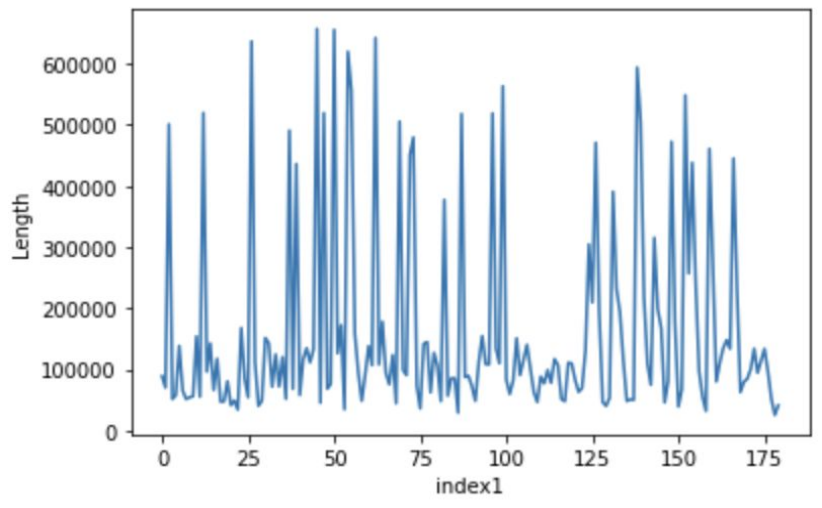
Foram gerados datasets de 15 minutos de duração total (900 segundos), contendo 30 fluxos anômalos e 150 fluxos normais, com 5 segundos de duração cada. O script também informa em quais intervalos de tempo os fluxos de anomalia foram executados para que se possa classificar o dataset posteriormente.

Definição do fluxo e geração dos datasets

Treino



Teste



Geração das estruturas probabilísticas

Tamanho do contador: 1000 posições;

Entrada do hash: IP de Origem, IP de Destino, Porta de Origem, Porta de Destino;

Função de Hash: MurmurHash3 32-bits + split em 1000;

Dado armazenado: Tamanho do pacote

Idade do contador: 5 segundos

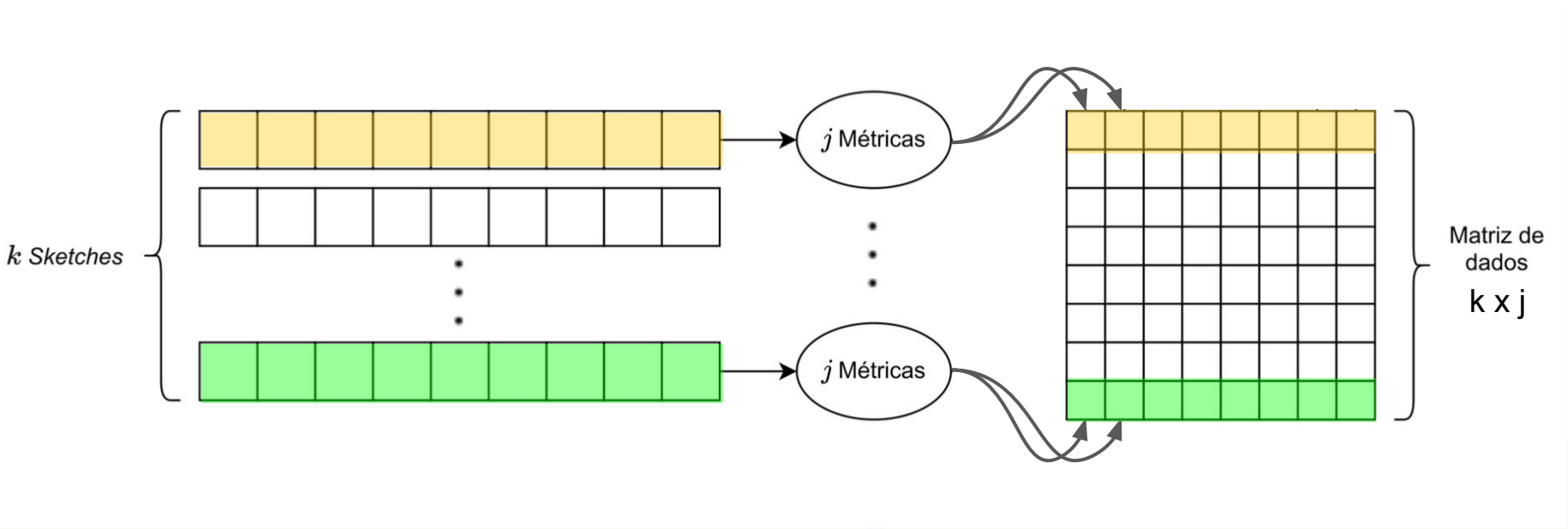
Extração de métricas

A partir de cada contador, foram extraídas as seguintes métricas:

`soma dos elementos`	`desvio padrão (σ)`
`número de não-zeros`	`mediana`
`maior valor`	`no. de elementos 2x maiores que σ`
`índice do maior valor`	`no. de elementos 3x maiores que σ`
`média`	

Cada contador representa um intervalo de tempo de 5 segundos (isto é, uma amostra) e cada métrica representa um atributo dessa amostra.

Extração de métricas



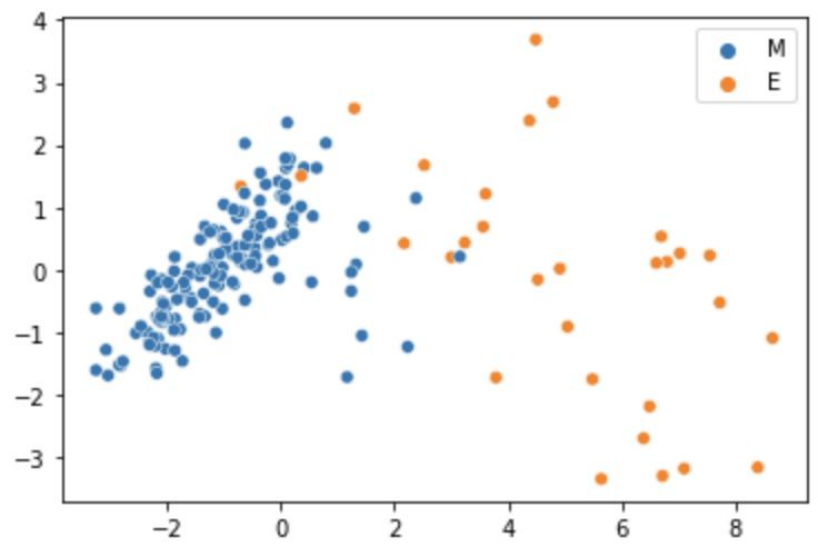
(PICOLI, 2022). Adaptado.

Extração de métricas

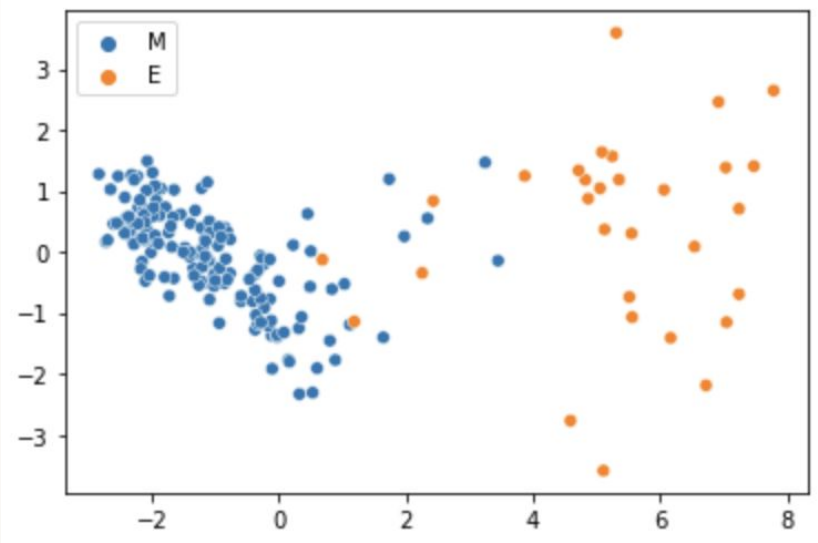
	Total	NonZeros	Max	MaxIdx	Mean	Desv	TwoTimes	TheeTimes	Mediana
0	49550	48	3753	33	1032.291667	993.910060	46	48	560.0
1	30600	22	3753	2	1390.909091	1312.424228	22	22	560.0
2	35850	29	4509	3	1236.206897	1516.848738	24	29	560.0
3	123650	92	8874	61	1344.021739	1607.024194	86	91	560.0
4	78750	56	5024	25	1406.250000	1542.317265	50	56	560.0
...	...	...	...	...	...	...	...	...	...
176	34180	39	3744	15	876.410256	856.008648	37	37	560.0
177	103890	76	4772	64	1366.973684	1431.654848	67	76	560.0
178	156080	104	5634	1	1500.769231	1579.300619	95	104	560.0
179	92180	69	5447	25	1335.942029	1405.038778	62	69	560.0
180	20840	18	4185	8	1157.777778	1225.215426	16	18	560.0

Visualização com PCA

Treino



Teste

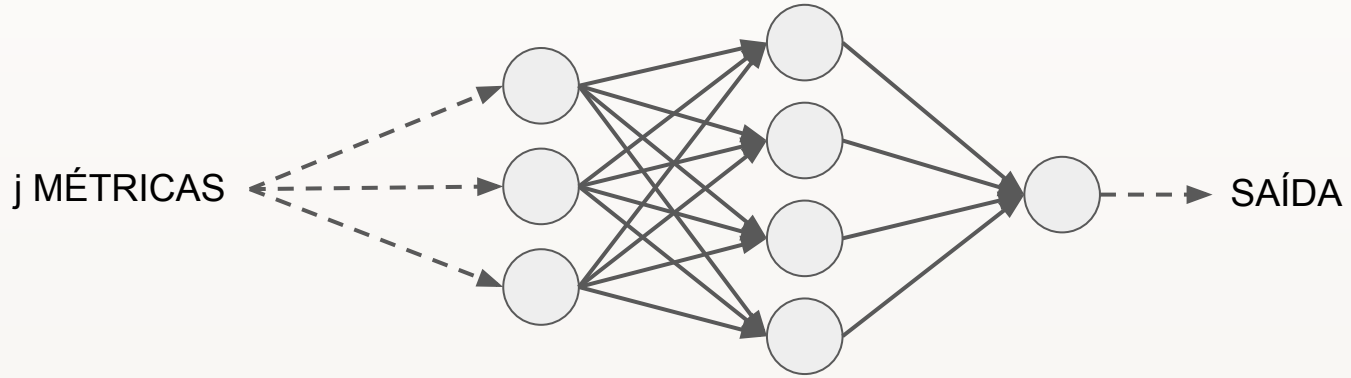


Variância retida: 76.38%, 13.94%, …

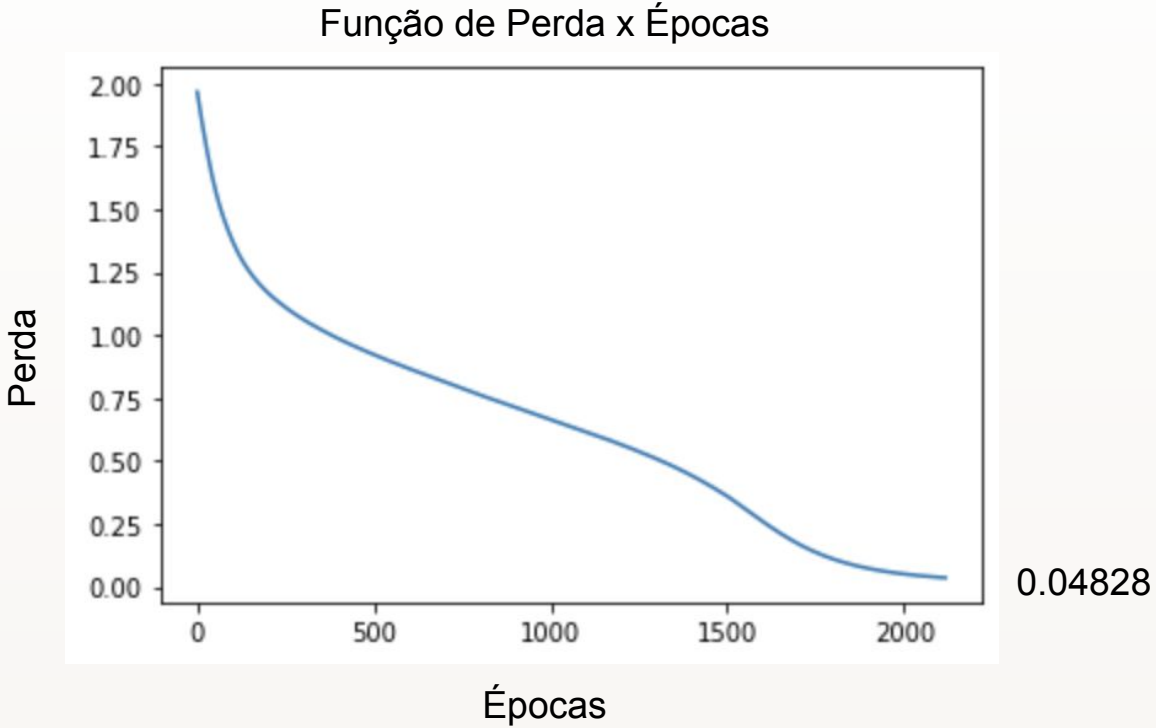
Classificação dos dados

Para a classificação dos dados, foram gerados 3 datasets: um de 10 minutos, para treino; um de 5 minutos para validação e um de 15 minutos para teste.

Como método, optamos por utilizar uma Rede Neural simples, com 3 neurônios na entrada, 4 na camada oculta e 1 na saída.



Classificação dos dados



0.04828

Resultado

97.79% de acurácia

Resultado

Comparando com outros métodos tradicionais:

	Rede Neural	Random Forest	AdaBoost	Naive Bayes
TN	**149**	148	150	145
TP	**28**	28	27	28
FP	**2**	3	1	6
FN	**2**	2	3	2
Acurácia	97.79%	97.23%	97.79%	95.58%

Obrigado!